

On Evaluating Video Object Segmentation Quality: A Perceptually Driven Objective Metric

Elisa Drelie Gelasca, *Member, IEEE*, and Touradj Ebrahimi, *Member, IEEE*

Abstract—The task of extracting objects in video sequences emerges in many applications such as object-based video coding (e.g., MPEG-4) and content-based video indexing and retrieval (e.g., MPEG-7). The MPEG-4 standard provides specifications for the coding of video objects, but does not address the problem of how to extract foreground objects in image sequences. Therefore, for specific applications, evaluating the quality of foreground/background segmentation results is necessary to allow for an appropriate selection of segmentation algorithms and for tuning their parameters for optimal performance. Many segmentation algorithms have been proposed along with a number of evaluation criteria. Nevertheless, formal psychophysical experiments evaluating the quality of different video foreground object segmentation results have not yet been conducted. In this paper, a generic framework for both subjective and objective segmentation quality evaluation is presented. An objective quality assessment method for segmentation evaluation is derived on the basis of perceptual factors through subjective experiments. The performance of the proposed method is shown on different state-of-the-art foreground/background segmentation algorithms and our method is compared to other objective methods which do not include perceptual factors. Moreover, on the basis of subjective results, weighting strategies are introduced into the proposed metric to meet the specificity of different segmentation applications e.g., video compression, video surveillance and mixed reality. Experimental results confirm the efficiency of the proposed approach.

Index Terms—Foreground/background extraction, mixed reality, objective evaluation, perceptual metric, psychophysical tests, segmentation, subjective quality assessment, video object, video object compression, video surveillance.

I. INTRODUCTION

UNSUPERVISED segmentation of digital images is a difficult and challenging task [1] with several key-applications in many fields: image classification, object recognition, etc. The performance of algorithms for subsequent image or video processing, compression, and indexing, to mention a few, often depends on a prior efficient image segmentation in which the *a priori* knowledge of the application is also integrated.

Manuscript received April 30, 2008; revised December 04, 2008. Current version published March 11, 2009. This work was supported by B.S. Manjunath and the National Science Foundation under Grant ITR-0331697. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lina Karam.

E. Drelie Gelasca is with the Center of BioImage Informatics, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: elisa.drelie@a3.epfl.ch).

T. Ebrahimi is with the Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSTSP.2009.2015067

Recent multimedia standards and trends in image and video representation [2] have increased the importance of adequately extracting (from the static background) moving foreground “objects” in video, in order to ensure efficient coding, manipulation, and identification [3]. For the remaining of the text, we will use the general term “segmentation” to refer to foreground/background segmentation and “foreground object” to refer to semantically meaningful disjoint regions.

Many segmentation algorithms have been proposed in the literature [4]–[9], as well as a number of evaluation criteria for segmentation quality assessment reviewed in Section II. The need for a quality metric arises from the fact that segmentation is an ill-posed problem: for the same image/video, the optimum segmentation can be different depending on the application.

Many researchers prefer to rely on qualitative human judgment for evaluation. However, subjective evaluation requires a large panel of human observers, resulting in a time-consuming and expensive process. Therefore, there is a need for an automatic objective methodology to allow for the appropriate selection of segmentation algorithms as well as to adjust their parameters for optimal performance.

In recent years, some objective methods for video object segmentation evaluation have been proposed, but no work has been performed on studying and characterizing the artifacts typically found in digital video object segmentation to derive a *perceptual* metric. A good understanding of the degree of annoyance of these artifacts and how they combine to produce the overall annoyance is an important step in the design of a reliable *perceptual objective quality metric* [10]. To this end, first a series of specifically designed psychophysical experiments has to be performed. The block diagram of the factors involved in deriving the perceptual objective metric is depicted in Fig. 1. The segmented video sequences can be thought of as being made of a combination of *ground truth* and *artifacts*. In this paper, we will use interchangeably the terms ground truth, reference or ideal segmentation. First, the mismatching regions are found by overlapping the ground truth to the segmentation under test and these regions are carefully classified and quantified in *objective errors*. In the block diagram, the ground truth link to the objective block is dotted, as ground truth may or may not be used to derive these features. If it is used, as in the scope of this paper, the method is called *discrepancy* or *reference method*. These objective errors are then combined in the overall quality measure by some mathematical relations to form the *objective metric*. The goal is to find the perceptual functions which link the objective metrics to the subjective (perceived) overall quality of the segmentation (mean opinion score, MOS). In the block diagram, it is described how the artifacts present in the segmented

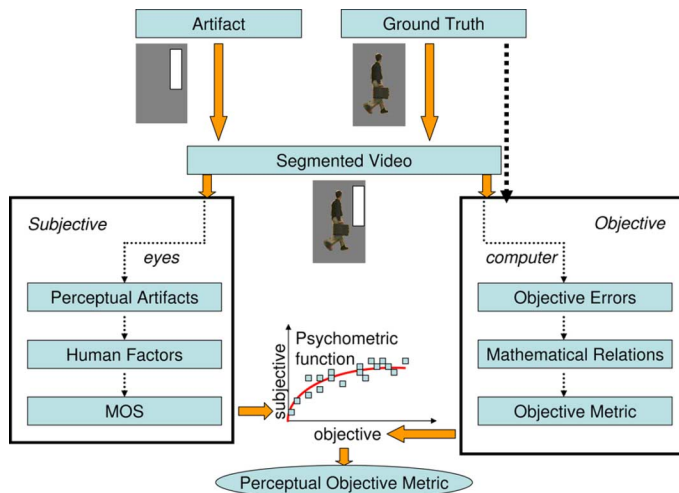


Fig. 1. Block diagram of the factors involved from subjective to objective video object segmentation quality evaluation.

video under test become “perceptual” when analyzed through the human visual system and how then the human visual perception (e.g., distortion in uniform regions are more visible than those in textured regions) come to play a fundamental role in the resulting MOS. *Psychometric functions* [11] are used to fit the relation between objective errors and MOS, since they better model the human perception. Hence, the graph plotting these objective and subjective quantities is related by a psychometric fitting curve that has to be determined to model the *perceptual objective metric* (as depicted in the bottom center of Fig. 1).

This paper aims to:

- 1) propose a formal framework to carry out psychophysical experiments in subjective video object segmentation evaluation;
- 2) derive a perceptual objective metric for segmentation evaluation on the basis of the subjective experiments on *synthetic* artifacts;
- 3) test the proposed metric on *real* segmentation algorithms and show its performance through subjective experiments;
- 4) show the good performance of the proposed perceptual objective metric with respect to state-of-the-art metrics which do not include perceptual factors;
- 5) find the tuning of the proposed quality metric parameters for different segmentation applications.

Preliminary results on the proposed metric have been previously published by the authors of this paper [12]–[16]. The main contribution of this work is the subjective and objective evaluation obtained for other segmentation applications taken into consideration. This paper not only gives an overall and complete overview of how to derive a perceptual metric for segmentation evaluation but also summarizes all the findings about subjective experiment protocol and segmentation artifacts annoyance. This work shows the performance of the proposed metric on new applications providing guidelines on which state-of-the-art segmentation algorithm (described in Section III) performs better for which “specific” video content/application.

The paper is structured as follows. First, a perceptual metric is built on synthetic artifacts. The novelty of the proposed ap-

proach consists of studying and characterizing the typical segmentation errors from a perceptual point of view. Different clusters of error pixels are perceptually classified according to the fact whether they do or do not modify the shape of the object. The goal of this step is to find a way to predict how people will judge the quality of segmentation without performing any subjective test. To achieve this goal, the following questions need to be addressed. 1) What method do people use to judge the segmentation quality? 2) Do people generally agree on the quality of a segmentation that is not trivial? 3) Does the expectation of quality affect ratings? An experiment should, if designed correctly, at least answer one or more of these questions. To this end, with the help of experts in psychophysical testing, we designed a series of psychophysical experiments in Section IV. This part of the work aims to develop standard methods for subjective evaluations of segmentation evaluation. In these experiments, we used test sequences with synthetic artifacts that look like *real* artifacts, but are simpler, purer, and easier to describe and control. In Section V, a perceptual objective metric is derived through subjective experiments conducted on synthetically generated artifacts. In Section VI an objective and subjective study of the annoyance generated by real artifacts introduced by typical video object segmentation algorithms is presented and the metric shows excellent performance for real artifacts. The objective and subjective study of the annoyance generated by real artifacts introduced by typical video object segmentation algorithms is presented both for generic framework and some specific applications: video compression, video surveillance and mixed reality. By “generic,” we imply that the final application of the segmentation under test has not yet been determined and the aim is to extract foreground objects as well as possible along the contours. In a generic scenario, the degree of annoyance is essential to evaluate segmentation quality but in specialized applications other parameters may be more suitable (e.g., in surveillance applications, correct detection is more important than the quality of the mask) as we will discuss. This paper also compares the performance of the proposed perceptual metric to state-of-the-art metrics. Conclusions are provided in Section VII.

II. OVERVIEW ON EVALUATION METHODS

The challenge of *subjectively* and *objectively* assessing the quality of segmentation has been investigated in the following different contexts in the literature: edge-based segmentation [17], region-based segmentation [18], and video object segmentation [19]–[25]. Nevertheless, there is no standardized procedure for subjective tests on any of these segmentation methods, nor any universally adopted objective metrics. In the literature (see Section II-A), subjective judgments are based on human intuition.

Subjective segmentation evaluation is necessary to study and to characterize the perception of different artifacts on the overall quality, but once this task has been accomplished successfully and an automatic procedure has been devised, systematic subjective evaluation can be avoided.

The automatic procedure is referred to as *objective evaluation method*. Quality metrics for objective evaluation of segmentation may judge either the segmentation algorithms

TABLE I
OBJECTIVE MEASURES USED IN EVALUATING IMAGE AND VIDEO OBJECT SEGMENTATION SYSTEMS

Criteria	Source
Positions of mis-segmented pixels	Cav. [15], Erdem [25], Villegas [22], [27]
Classes of mis-segmented pixels	Cav. [15], Villegas [22], MPEG [23]
Number of objects	Correia [20], Nascimento [28]
Shape changes	Erdem [25], Correia [20], Mech [24]
Temporal stability	Villegas [22], [27], MPEG [23], Erdem [25], Cav. [15]
Temporal drift	Villegas [22], [27]

TABLE II
OBJECTIVE MEASURES USED IN EVALUATING VIDEO TRACKING SYSTEMS

Criteria	Source
False Alarm	Ellis [29], Nascimento [28], Oliveira [30], Oberti [31]
Misdetetection	Ellis [29], Nascimento [28], Oliveira [30], Oberti [31]
Split and/or Merge	Ellis [29], Nascimento [28], Oberti [31]
Area Matching	Ellis [29], Nascimento [28]
Occlusion management	Ellis [29]
Center of gravity	Ellis [29], Senior [32]

or their segmentation results. These are referred to as analytical or empirical methods, respectively [26]. *Empirical methods* do not evaluate the segmentation algorithms directly, but indirectly through their results. Empirical methods are divided into *empirical discrepancy* (or *reference*) metrics when the segmentation result is compared to an ideally segmented “reference” mask (ground truth), and *empirical goodness* (or *no-reference*) metrics when the quality of the result is based on intuitive measures of goodness such as color uniformity. The main disadvantage of such an approach is that the goodness metrics are at best heuristic, and may exhibit strong bias toward a particular algorithm. In other words, it is extremely difficult to define segmentation goodness criteria that are valid in general for any application. For example, the intra-region gray-level uniformity goodness metric will cause poor evaluation for any segmentation algorithm that forms regions of uniform texture. On the other hand, a given ground truth defines the application and the requirements of the segmentation algorithm *a priori*, and it allows for developing a more flexible metric. For this reason, we have chosen to implement a discrepancy method, which makes use of the ground-truth. State-of-the-art discrepancy methods are reviewed in Sections II-B and II-C and summarized in Tables I and II.

A. Subjective Evaluation

A set of general guidelines for segmentation quality assessment has been proposed in the COST211/quat European project [19]. These guidelines concern only how the typical display configuration should appear (see [20]), but do not specify how the test should be carried out (e.g., experimental methodology such as type of questions to observers, displayed segmentation, etc.). This framework proposes to show people four video sequences at the same time and it does not specify how long they should be. Here, we performed informal tests and observed that using this display configuration, four video sequences (5–10 s), at the same time, were too many because subjects can concentrate only on one of them. Moreover, this layout also shows the original sequence without any segmentation, which we believe is not essential since the subject, once he/she has learned the task, forms

his/her own *implicit* segmentation and no longer looks at the original nor at the reference segmentation.

In [21], some criteria related to the computational complexity of the segmentation system are defined together with a number of questions to investigate subjectively the video object segmentation quality for surveillance applications. For each video sequence, the subject can see the original video sequence as many times as necessary. Then, the segmented video is presented only once and the subject has to answer four evaluation criteria questions (such as “how well have important moving objects been individually identified?”, or “how well have boundaries been provided?”).

In informal tests, we tried to combine different questions to describe the aspects of segmentation quality. However, we noticed that in this case subjects had to perform a sort of *memory test* given the large number of questions asked after the video was played back. The capacity of a test subject to reliably assess several elements of a video is limited. The memory of a video fades after time.

For all the above described reasons, a new subjective evaluation methodology is proposed in Section IV, in which only one question is asked after the video is played back and one video sequence is shown during the test.

B. Video Object Segmentation Evaluation

In this section, the advantages and disadvantages of state-of-the-art objective methods for segmentation evaluation [15], [20], [22], [20], [23], [24], [22], [28] are presented, none of which includes the characterization of artifact perception in their models. To evaluate a segmented video by discrepancy methods, Erdem and Sankur [25] combined three empirical discrepancy measures into an overall quality segmentation evaluation: *misclassification penalty*, *shape penalty*, and *motion penalty*. In [20], first the individual segmentation quality is measured by four spatial accuracy criteria: *shape fidelity*, *geometrical fidelity*, *edge and statistical content similarity* and two temporal criteria: *temporal perceptual information* and *criticality*. Second, the similarity factor between the reference and the resulting segmentation is computed. Furthermore, the multiple-object case was addressed by using the criteria of application-dependent “*object relevance*” to provide the weights for the quality metric of each object. Finally, they combined all these three measures into an overall segmentation quality evaluation.

Another way to approach the segmentation evaluation problem is to consider it as a particular case of shape similarity as proposed in [24] for video object segmentation. In this method, the evaluation of the spatial accuracy and the temporal

coherence is based on the mean and standard deviation of the 2-D shape estimation errors. In preliminary work, we proposed to evaluate the quality of a segmented object through spatial and temporal accuracy joined to yield a combined metric [15]. This work was based on two other discrepancy methods [23], [27] described below, but did not include any perceptual factor. A summary table of these state-of-the-art methods, categorized according to their evaluation criteria, is presented in Table I.

During the standardization work of ISO/MPEG-4, within the core experiments on automatic segmentation of moving objects, it became necessary to compare the results of different object segmentation algorithms by subjective evaluation as well as by objective evaluation. The proposal for objective evaluation [23] agreed upon by the working group, uses a ground truth. The *MPEG error measure* metric is adopted by the research community because of its simplicity. A refinement of this metric, *weighted quality metric*, has been proposed by Villegas *et al.* [22], [27]. Since these are the two metrics most commonly adopted by the research community when evaluating a segmentation algorithm, we will use them for comparison to our proposed metric. Aside from these two metrics, the metric presented in Section II-C is usually used for video tracking evaluation and it is more relevant when the segmentation algorithm under test is used in a video surveillance application as shown in Section VI. For this reason, the metric described in Section II-C has also been chosen for comparison to the new metric proposed in this paper.

1) *MPEG Evaluation Criteria*: A moving object can be represented by a binary mask, called an *object mask*, where a pixel has an object-label if it is inside the object and a background-label if it is outside the object. The objective evaluation approach used in the ISO/MPEG-4 core-experiment has two objective criteria: the *spatial accuracy* and the *temporal coherence*. Spatial accuracy, Sqm , is estimated through the amount of error pixels in the object mask (both false positive and false negative pixels) in the resulting mask deviating from the ideal mask.

Temporal coherence is estimated by the difference of the spatial accuracy between the mask M at the current and previous frame k

$$Tqm_M(k) = Sqm(k) - Sqm(k-1). \quad (1)$$

The two evaluation criteria can be combined in a single *MPEG error measure*, through the sum

$$\text{MPEG} = \frac{1}{K} \sum_k (Sqm(k) + Tqm_M(k)). \quad (2)$$

In this metric, the perceptual difference of different classes of errors, false positive and false negative, is not considered and they are all treated the same. In fact, different kinds of errors should be combined in the metric in correct proportions to match evaluation results produced by human observers.

2) *Weighted Evaluation Criteria*: Within the project COST 211 [19], the above approach has been further developed by Villegas and Marichal [22], [27]. For the evaluation of the spatial accuracy, as opposed to the previous method, two classes of pixels are distinguished: those which have an object-label in the

resulting object mask, but not in the reference mask (false positive) and vice versa (false negative), and they are weighted differently. Furthermore, their metric takes into account the impact of these two classes on the spatial accuracy, that is, the evaluation worsens with pixel distance d to the reference object contour. The spatial accuracy qms is normalized by the sum of the areas of reference objects as follows:

$$\begin{aligned} qms(k) &= \frac{qms^+(k) + qms^-(k)}{\sum_{i=1}^{N_R} R_i(k)} \\ &= \frac{\sum_{d=1}^{D_M^+} w_+(d) \cdot |\mathcal{P}_d(k)| + \sum_{d=1}^{D_M^-} w_-(d) \cdot |\mathcal{N}_d(k)|}{\sum_{i=1}^{N_R} R_i(k)} \end{aligned} \quad (3)$$

where D_M^+ and D_M^- are the biggest distance d for, respectively, false positives and false negatives; N_R is the total number of foreground disjoint regions in the reference R ; $\sum_{i=1}^{N_R} R_i(k)$ is the sum of the area of all the objects i in the reference; $\mathcal{P}_d(k)$ and $\mathcal{N}_d(k)$ are positive and negative pixels, respectively; $w_+(d)$ and $w_-(d)$ are the weights for positives and negatives respectively, expressed as

$$w_+(d) = b_1 + \frac{b_2}{d + b_3}, w_-(d) = f_S \cdot d \quad (4)$$

where the parameters b_i and f_S are chosen empirically [22]: $b_1 = 20$, $b_2 = -178.125$, $b_3 = 9.375$ and $f_S = 2$. These functions represent the fact that the weights for false negative pixels increase linearly and they are larger than those for false positives at the same distance from the border of the object. However, as we move away from the border, missing parts of objects become more important than added background.

Two criteria are used for estimating temporal coherence: the temporal stability $qmt(k)$ and the temporal drift $qmd(k)$ of the mask. First, the variation of spatial accuracy criterion between successive frames is investigated as follows. The temporal stability is equal to the normalized sum of the differences of the spatial accuracy for two consecutive frames for false positive and false negative pixels

$$qmt(k) = \frac{qms^+(k, k-1) + qms^-(k, k-1)}{\sum_{i=1}^{N_R} R_i(k)}. \quad (5)$$

where $qms^*(k, k-1) = |qms^*(k) - qms^*(k-1)|$.

Second, the displacement of the gravity center \vec{G} of the resulting object and the reference object mask is computed for successive frames to estimate possible *drifts* of the object mask $\vec{qmd}(k)$

$$\vec{qmd}(k) = [\vec{G}_E(k) - \vec{G}_R(k)] - [\vec{G}_E(k-1) - \vec{G}_R(k-1)] \quad (6)$$

that is displacement from time $(k-1)$ to time (k) of the centers of gravity \vec{G} of the estimated E and reference R masks. The value of drift is the norm of the displacement vector divided by the sum of the reference object bounding boxes (area)

$$qmd(k) = \frac{\|\vec{qmd}(k)\|}{\frac{1}{N_R} \sum_{i=1}^{N_R} \text{BB}_i^{x,y}(k)} \quad (7)$$

where $\text{BB}_i^{x,y}(k)$ is the bounding box area of the object i in the reference mask R at time k . The authors proposed to define a

single quality value by linearly combining all the three presented measures as the **weighted quality metric** $wqm(k)$

$$wqm(k) = w_1 \cdot qms(k) + w_2 \cdot qmt(k) + w_3 \cdot qmd(k), wqm = \frac{1}{K} \sum_k wqm(k). \quad (8)$$

The values of the weights w_i are very much application dependent. If no application is specified, all three weights can be assumed equal to $(1)/(3)$.

In this method, the perceptual difference between two kinds of errors is taken into account. The drawback is that the weighting functions defined in (4), that should be “perceptual” weights of the evaluation criteria, are defined by means of empirical tests. These empirical tests are not generally sufficient. As well in all other proposed evaluation criteria in the literature, the relevance and the corresponding weight of different kinds of errors should be supported by formal subjective experiments performed under clear and well defined specifications.

C. Video Object Tracking Evaluation

Recently, a number of measures have been proposed for video object tracking evaluation. Since we are interested in how the object is segmented and the evaluation of tracking raises different problems briefly discussed in this section, the reader is introduced to fora such as PETS [33] and CAVIAR [34] for a complete overview on that issue.

In the following, we will refer to some representative works [29]–[32] that can be found in the literature and specifically to Nascimento and Marques’s metric [28] that can be applied also to a more general object segmentation evaluation case. Table II shows all the state-of-the-art methods grouped by discrepancy measure.

Standard measures used in communication theory such as misdetection rate, false alarm rate, and receiver operating characteristics (ROC) are used in [30], [31]. A ROC curve is generated by computing pairs (P_d, P_f) , where P_d is the probability of correct signal detection and P_f is the false alarm probability. For example, Oberti *et al.* [31] computed the false-alarm (P_f) and the misdetection probabilities $(1 - P_d)$ on the basis of discrepancies between the resulting objects and matching area (false alarm) or between the reference area and the matching one (misdetection). The global performance curve summarizing the curves obtained under different working conditions is obtained by imposing an operating condition ($P_f = 1 - P_d$) and by plotting the corresponding values against different values of the variable of interest (scene complexity, distance of objects from sensors).

In [30], a specific parameter of the tracking algorithm is varied and the false alarm/detection and split/merge rates are plotted against it. Senior *et al.* [32] employed the trajectories of the centroids of tracked objects and their velocities to evaluate their discrepancy measures.

An interesting framework for tracking performance evaluation uses pseudosynthetic video [29]. Isolated ground truth tracks are automatically selected from the PETS2001 dataset, according to three criteria: path, color, and shape coherence (in order to remove tracks of poor quality). Pseudosynthetic video

sequences are generated by adding more ground truth tracks and the complex object interactions are controlled by the tuning of perceptual parameters. The metrics used are similar to those in the previously described works: tracker detection rate, false alarm rate, track detection rate, occlusion success rate, etc.

However, these approaches have several limitations. As already mentioned, object detection cannot be considered as a simple binary detection problem. Several types of error should be considered; misdetection and false alarms alone are not sufficient. For example, the proposed test in [32] is based on employing the centroid and areas of rectangular regions, but practical algorithms have to segment the image into background and foreground and should not classify rectangular regions selected by the user.

To overcome these limitations Nascimento and Marques [28] used several simple discrepancy metrics to classify the errors into region splitting, merging or split-merge, detection failures, and false alarms. In their scenario, the most important thing is that all the objects have to be detected and tracked along time. Object matching is performed by computing a binary correspondence matrix between the segmented and the ground truth images. The advantage of the method is that ambiguous segmentations are considered (e.g., it is not always possible to know if two close objects correspond to a single group or a pair of disjoint regions: both interpretations are adopted in such cases). In fact, by analyzing this correspondence matrix, the following measures are computed: Correct Detection (C_D): the detected region matches one and only one region; False Alarm (F_A): the detected region has no correspondence; Detection Failure (D_F): the test region has no correspondence; Merge Region (M): the detected region is associated to several test regions; Split Region (S): the test region is associated to several detected regions; Split-Merge Region (S_M): when the conditions M and S simultaneously occur.

The normalized measures are obtained by normalizing the amount of F_A by the number of objects in the segmentation, N_C , all the others by the number of objects in the reference, N_R , and by multiplying the obtained numbers by 100. The **object matching quality metric** at frame k , $mqm(k)$, is finally given by

$$mqm(k) = w_1 \cdot \frac{C_D(k)}{N_R} + w_2 \cdot \frac{F_A(k)}{N_C} + w_3 \cdot \frac{D_F(k)}{N_R} + w_4 \cdot \frac{M(k)}{N_R} + w_5 \cdot \frac{S(k)}{N_R} + w_6 \cdot \frac{S_M(k)}{N_R} \quad (9)$$

where w_i are the weights for the different discrepancy metrics. mqm is the sum of $mqm(k)$ normalized over all frames. It is evident that this metric is able to describe quantitatively the correct number of detected objects and their correspondence with the ground truth only, while the metrics described in the previous sections are able to monitor intrinsic properties of the segmented objects such as shape irregularities and temporal instability of the mask along time.

III. SEGMENTATION ALGORITHMS

In our experiments, we chose seven static background segmentation methods. The approaches of the selected representative algorithms differ in using various features such as color,

luminance, edge, motion, and combinations of them. A quick overview of the principles on which each technique is based is reported. The background frame is available and used to extract the foreground objects in most of these techniques. For further details the reader is invited to refer to each appropriate paper. Tuning of parameters has been done on several video sequences and the best parameters for each algorithm were tuned according to visual inspection.

Image Differencing is based on basic background subtraction in which grayscale images are used and an absolute differencing with the background and current frame is applied. The segmentation results depend only on the threshold method used for obtaining the binary mask (foreground/background). The Otsu thresholding method is used [35]. The segmentation results differ very much since the threshold value is sensitive to environmental conditions, e.g., similar colors, illumination changes.

Kim's [4] approach is based on gray level images and applies the Canny edge operator to the current, background, and successive frames. The shape information for moving objects is obtained by the edge map of difference frames that is used together with the background edge map for selecting the relevant edges in the current frame. Finally, the object mask is achieved by filling the boundaries obtained by the previous edge results with connecting the first and second occurred edge pixels for each vertical and horizontal line, respectively.

Horprasert [6] assumes that the luminance and chrominance has to be separated on the RGB color space by generating a new color model. In that, there is an expected chromaticity line in which the pixel value should be kept. The expected chromaticity is obtained by the arithmetic means for each pixel of the RGB values calculated over a number of background images. The distortion from this line is given as both chromaticity and brightness distortion being generated by standard deviation. With these distortions several thresholds are determined to classify the pixel to one of the following types: *original background*, *shadow*, *highlighted background*, and *foreground*.

François and Medioni's [5] technique operates assuming that in the background only very slow global changes can occur and further, the color values of each pixel build a spherical cluster in the color space. With these assumptions, a background model based on a Gaussian distribution is generated by considering the mean value and standard deviation for each pixel. In the system, the HSV color space is used instead of the RGB. The current image is subtracted from the mean value model and the resulting difference values of each pixel give the information of classifying to either foreground or background with regard to the standard deviation model. Moreover, an update of the background model is also given.

Shen [7] uses the RGB color space and the system can be represented in two steps. One of them is the block for generation of fuzzy classification and the other one is the block for elimination of falsely detected segmentation regions. The fuzzy classification is applied to take into account the mobility of pixels precisely instead of the so-called binary classification. Thus, in the fuzzy block a difference image is generated for each RGB color space component. For every channel result a corresponding threshold is determined by using a unimodal thresh-

olding method for considering the fuzzy set of mobile pixels. Then these thresholds avail to generate fuzzy images which are then combined to one final fuzzy image. Subsequently, a preliminary mask is achieved by thresholding. It describes all detected mobile pixels in all appearances. To overcome the problems of illumination changes and since there is no sudden adaptive update of the background, a combination of temporal information and the mentioned above fuzzy color classification is given. The temporal information is achieved by the OR operation of the image differencing of successive frames and the last resulting mask. This output is combined with the preliminary mask of the fuzzy classification block.

Jabri's [8] system uses both information: the RGB pixel color values and the edge. The background model is trained in both mentioned parts by calculating the mean and standard deviation for each pixel of any color channel. With background subtraction of the incoming current image on each channel, confidence maps are generated for both information color and edge. After that, a combination of the two maps are utilized by taking its maximum values. At last, this output goes through a hysteresis thresholding for binarization.

McKenna *et al.* [9] also uses color and edge information to model the background. Instead of the RGB color space the normalized RGB space (rgb) is used. The models are generated separately for each channel. The incoming frame is classified separately and a combination of both classification results gives the final segmentation mask.

IV. SUBJECTIVE EVALUATION

The proposed subjective experiment methodology corresponds to the sequential five-step procedure described in detail in [16].

- 1) *oral instructions*: the subject is made familiar with the task of segmentation and with the original video sequences;
- 2) *training*: the subject is introduced to original video, reference segmentations (best quality case) and segmentations with the most annoying artifacts where subjects are asked to mentally attribute 100 to the most annoying one;
- 3) *practice trials*: subjects' responses between 0 and 100 are collected on a small subset of test sequences;
- 4) *experimental trials*: the test is performed on the complete set of sequences;
- 5) *interview*: qualitative descriptions of the perceived artifacts collected after the test is carried out in order to improve the design of future experiments, subjects do not have access to segmentation results.

The standard protocol [36] used for the subjective assessment is Single Stimulus with a continuous (0–100) scale. Standard methods [37] are used to analyze and to screen the judgments provided by the test subjects. The data is first processed by calculating the MOS. Second, outliers are rejected by a screening standard procedure [37]. In this context, the MOS is called Mean Annoyance Value (MAV) since in this case the subjective scores correspond to "annoyance" scores.

The reference segmentations used in the subjective experiments were manually obtained by the authors or by the MPEG group [3]. It is interesting to note that for some reference segmentations the values for MAVs corresponding to them are not

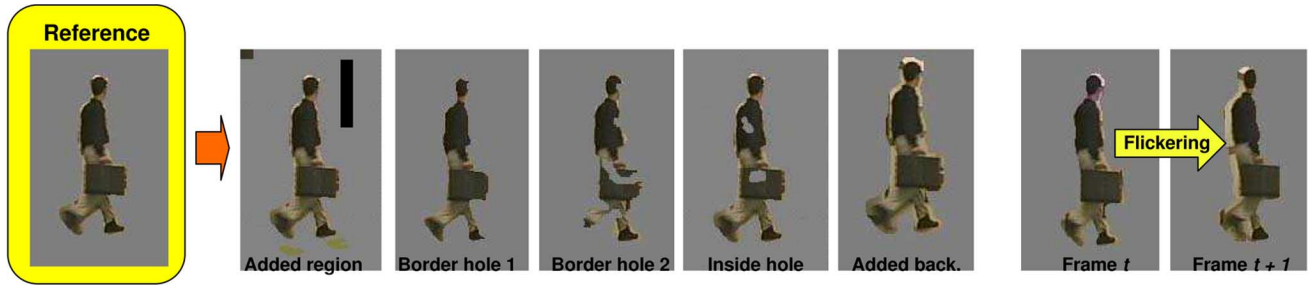


Fig. 2. Reference (ground truth) segmentation with common segmentation artifacts: added region, two examples of border hole, inside hole, added background, and flickering (any spatial artifact varying over consecutive frames).

zero, indicating that subjects report that these segmentations contained some type of annoyance level different from zero. This is due to the fact that the reference segmentation manually obtained is not perfect and few pixels can be erroneously segmented along the object contours. The subjects are able to mentally form the ideal segmentation and to detect and judge such small imperfections.

The test group was composed of 35 subjects. The subjects were asked one question after each segmented video sequence was presented; “How annoying was the defect relative to the worst example in the sample video sequences?” the subject was instructed to enter a numerical value greater than 0. The value 100 was to be assigned to artifacts as annoying as the most annoying artifacts in the sample video sequences in the training phase. The subjects were then told that different artifacts would appear combined or alone and they should rate the overall annoyance in both cases. As a result of the survey given to the subjects, *five* different clusters of errors were recognized during the subjective experiments as typically provided by the most common segmentation algorithms (described in the next section). These five artifacts are depicted in Fig. 2 along with the reference (ideal) segmentation. **Added region** is the over-segmented part of background disjoint from the correctly segmented objects that does not form any semantically meaningful region. **Added background** is the over-segmented part of background attached to the correctly segmented object that makes the object larger. **Inside holes** are under-segmented parts completely contained inside the objects that are visible through the object parts of the background. **Border holes** are under-segmented parts directly attached on the border of the object and that make the object thinner. **Flickering** is the temporal variation of any of the above described artifacts that makes the object suddenly changes its shape or meaningless regions to appear and disappear in the segmentation.

The three original sequences used in this experiment are “Group,” “Hall monitor,” and “Highway” [see Fig. 3(a)–(c)]. The seven segmentation algorithms described in the previous section have been applied to each original video sequence. Since the quality of segmentation is strongly connected to its application (e.g., compression, mixed reality or video surveillance) four different *frameworks* were considered in our subjective experiments. Both *general* and three *application* dependent segmentation frameworks were considered in the subjective evaluation. A total number of 96 sequences were

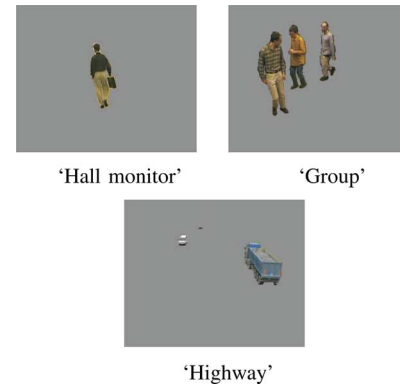


Fig. 3. Sample frames of the reference segmentation for the general framework.

generated: 24 test segmented sequences (3 original \times 7 segmentations plus 3 references) \times 4 frameworks. In the general scenario and with synthetic artifacts, the textured video objects have been overlapped on a uniform gray background ($Y = 127, U = 127, V = 127$) to less affect the human viewer according to the opinion of psychophysical experts.

In order to assess if a segmentation is good in a general scenario, viewers were asked to mentally compare the results of the segmentation at hand with the ideal (reference) segmentation (shown in Fig. 3) and formulate their judgments. Studying how subjective quality scores change in relation to the specific segmentation tasks provides a lot of interesting insights in developing evaluation metrics. In the following, a possible application scenario is described and the subjective results providing general guidelines for the development of segmentation algorithms are presented.

A. Scenario Dependent Evaluation

The expected segmentation quality for a given application can often be translated into requirements related to the shape precision and the temporal coherence of the objects to be produced by the segmentation algorithm. Video sequences segmented with high quality should be composed of objects with precisely defined contours, having a perfectly consistent partition along time. A large number of video segmentation applications can be considered and typically they have different requirements. The setting up of a subjective experiment differs for each application. Therefore, our experiments were focused on four types of scenarios for segmented objects: general,



Fig. 4. Sample frames for video coding segmentation applications “Hall monitor,” “Group,” and “Highway” with a zoom image of the background (for a better visualization of the compression artifacts it is suggested to print in color).

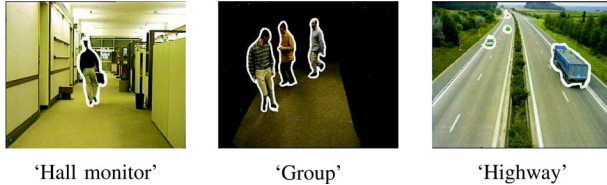


Fig. 5. Sample frames for video surveillance segmentation application “Group,” “Hall monitor,” and “Highway.”

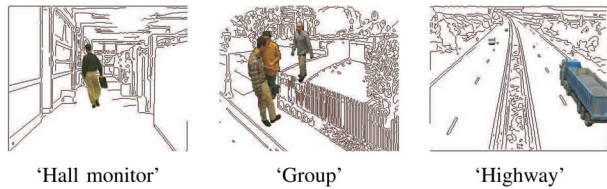


Fig. 6. Sample frames for video augmented reality segmentation application “Group,” “Hall monitor,” and “Highway.”

video compression, video surveillance, and mixed reality (see Figs. 3–6). For details on the scripts with the transcription of the verbal instructions given to the subjects, the reader is referred to [16].

For the **general** framework, we mean that the segmentation under test does not yet have a targeted application and it has to be evaluated in general, according to general goodness criteria of the segmented contour of foreground objects. Subjects are asked to mentally compare the results of the segmentation to the reference and provide an evaluation for the general purpose of effectively extracting the foreground objects according to the amount of added background, added region, missing regions, border, and inside holes. The extracted foreground objects are overlapped to a gray uniform background (see Fig. 3) so as not to influence the perception of artifacts according to the opinion of psychophysical experts in the Psychology department at University of California, Santa Barbara. In this scenario, subjects are instructed to judge the most annoying perceived artifact not according to a specific application but to how well the foreground object is extracted in general. As we will see in the next section, according to interviews conducted following the experiments, the most annoying segmented videos are those that contain the largest number of artifact types simultaneously and artifacts that are largest in size.

In **video compression**, segmentation can be used to improve the coding performance over a low-bandwidth channel. One of the functionalities of the MPEG-4 coding scheme is to support the compression of the background separately from the foreground objects so that more bits can be dedicated to the compression of the meaningful objects. Since we are only inter-

TABLE III
DESCRIPTION OF SEGMENTATION ALGORITHMS ARTIFACTS AND THEIR PERCEIVED STRENGTHS GATHERED IN THE INTERVIEW STAGE

Algorithm	Artifacts	Strength
Shen	added background border holes	low low
Jabri	added regions added background	medium low
Horprasert	border holes	medium
François	added background	high
McKenna	inside holes border holes flickering	medium medium medium
Image Differencing	inside holes border holes flickering	high high medium
Kim	added regions added background flickering	high high high

ested in studying the perception of segmentation artifacts, distortions due to compression were not included in the segmented foreground objects but were in the background. Thus, the segmented video objects were not compressed. In this way, the compressed background and uncompressed foreground can be transmitted only once, and the video objects corresponding to the foreground (moving objects) could be transmitted and added over it so as to update the scene. The Microsoft MPEG-4 implementation (Microsoft’s MPEG VM software encoder and decoder¹) was used in the experiments. A sample of a compressed background test sequence is shown in Fig. 4. All the quantization parameters Q for the background coding were chosen to be equal to 10. Subjects were instructed as to the video compression principles and asked to only judge the foreground object segmentation quality. Video compression is a typical case where knowledge of the specific application can be used to tune the parameters of the evaluation metric: undetected object’s parts will have a bigger impact on the overall annoyance than over-segmentation of the detected objects (see Section VI). In fact, the parts of the object that are undetected will be compressed as erroneously considered parts of the background. By means of subjective tests, we proved the hypothesis that *border holes* are the most annoying artifacts for this specific application. Therefore, we found that McKenna and Image Differencing algorithms that introduce a lot of *border holes* are not suitable for compression applications (see Fig. 7) as we will discuss in Section VI.

Video Surveillance. For a specific application such as video surveillance, a different protocol is needed to carry out subjective experiments. Subjects are instructed about miss rate and false alarm errors and the importance of a good segmentation

¹Version: FDAMI 2-3-001213, integrator: Simon Winder, Microsoft Corp.

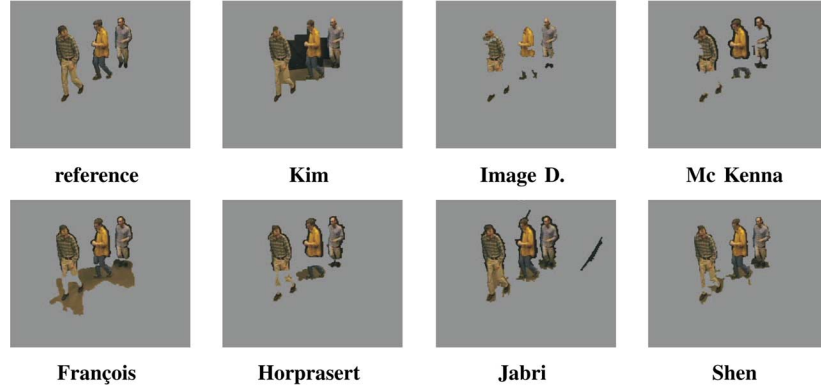


Fig. 7. Sample frames for the reference and some segmentation results of the tested video sequence “Group” (frame #100).

TABLE IV
MAV VALUES OBTAINED FOR EACH SEGMENTATION ALGORITHM FOR ALL THE TEST VIDEO SEQUENCES
IN GENERIC, COMPRESSION, SURVEILLANCE, AND MIXED REALITY FRAMEWORKS

Alg.	‘Group’				‘Hall monitor’				‘Highway’				MAV			
	Gen.	Cmpr.	Mix.	Sur.	Gen.	Cmpr.	Mix.	Sur.	Gen.	Cmpr.	Mix.	Sur.	Gen.	Cmpr.	Mix.	Sur.
ref.	8.77	11.20	12.54	13.45	26.74	15.51	12.60	17.84	15.31	14.45	8.03	10.02	16.94	12.5	13.20	13.77
Jab.	57.46	22.63	51.71	32.07	40.37	10.60	44.54	34.98	37.94	49.82	39.00	43.24	42.25	19.41	48.69	36.76
Shen	57.83	33.94	57.70	31.05	55.26	60.71	62.22	29.50	54.26	53.57	33.60	41.67	55.78	50.26	57.79	33.66
Hor.	69.94	48.63	72.80	65.98	57.57	20.17	57.14	65.50	32.06	42.62	34.77	29.08	53.19	29.8	57.52	53.52
Fran.	68.57	39.57	76.08	45.23	61.43	66.71	77.42	68.93	30.20	45.28	40.66	47.87	53.40	46.58	66.26	54.01
McK.	83.36	76.43	84.77	85.72	56.86	71.37	74.65	75.38	54.26	71.57	68.62	34.43	68.82	73.12	76.01	65.15
Im.D	99.74	90.00	95.08	73.37	60.00	48.40	57.68	43.41	67.54	75.34	79.42	80.15	75.76	71.24	77.40	65.64
Kim	72.00	40.00	73.00	56.78	86.89	52.00	82.54	78.43	71.14	45.51	84.51	65.78	76.67	45.8	80.01	66.99

in relation to these errors in order to detect intruders in indoor scenes or monitoring traffic in highways. In order to evaluate different segmentation algorithms in the context of a video surveillance application, the segmentation results and the reference segmentation have been used to produce test video sequences where the object boundaries detected by the segmentation algorithm have been underlined on the original video sequence by a colored contour as depicted in Fig. 5. In this scenario, *added regions* seem to be more critical than other artifacts by the subjects since they may provoke a false alarm as the artifacts introduced by Kim’s algorithm (see Fig. 7) discussed in the result Section VI.

Mixed Reality. The goal of video manipulation is to put together video objects from different sources in order to create new video content. In particular, in the *mixed reality* application [38] considered here, video segmentation serves to extract real objects that are then inserted in a virtual background. One of the possible applications is to create narrative spaces and interactive games and stories [39]. In order to evaluate different segmentation results in a mixed reality scenario, we created a virtual background for each original sequence: we extracted the contour of the background image to recall a virtual background in black and white as in comics scenarios. For the test sequence “Group” we applied a virtual background created in the context of the European Project art.live [39] processed the same way to extract only the contours. Fig. 6 shows sample frames. In this case, *inside holes* seem to be the most annoying since they create an annoying virtual background visible through the holes in the foreground objects. For example McKenna and François algorithms will be judged poorly because of the hole artifacts they

TABLE V
F VALUES TO TEST IF DIFFERENT FITTING CURVES ARE NEEDED TO DESCRIBE
THE PERCEIVED ANNOYANCE FOR DIFFERENT SHAPES AND POSITIONS
OF ADDED REGIONS AND HOLES

Artifact model	F_c (critical)	F (value)	$p(F < F_c)$
added region shape	$F(2,68)=3.13$	1.43	0.24
added region position	$F(4,66)=2.51$	0.64	0.63
inside hole position	$F(2,28)=3.34$	0.13	0.87
hole distinction	$F(2,44)=3.21$	5.01	0.01

largely introduce as presented in the experiments described in Section VI.

B. Subjective Results

From the data gathered, we obtained an overall opinion by interviewing the subjects and calculated the MAV of each test sequence. Table III shows the subjective opinion ranking gathered during the *interview stage* of the subjective experiment for the general framework. This table reports the tested algorithms from the least to the most annoying and a brief description of the artifacts that are typically introduced as a result of the surveys given to the subjects. According to the subjective opinion, the most annoying video segmentations were those which presented the largest number of artifact, the most types of artifacts simultaneously and the most flickering. This qualitative description helped us to better understand the quantitative values obtained in the experimental trials presented in Table IV.

Table IV shows the MAV values, gathered in the *experimental trials*, for all video and algorithms, along with the different sce-

TABLE VI
CORRELATION COEFFICIENTS BETWEEN THE OBJECTIVE METRICS AND SUBJECTIVE RESULTS (MAV VALUES) FOR ALL THE TEST VIDEO SEQUENCES IN GENERIC AND SPECIFIC APPLICATION FRAMEWORKS. PST METRIC PARAMETERS: $a = 2.86, b = 4.50, c = 4.77, d = 5.82$

Metric	'Generic'		'Compression'		'Mixed Reality'		Surveillance	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
wqm	0.69	0.71	0.37	0.32	0.74	0.65	0.79	0.77
mqm	0.53	0.44	0.50	0.47	0.67	0.55	0.72	0.65
MPEG	0.73	0.67	0.49	0.41	0.78	0.68	0.83	0.80
PST	0.86	0.79	0.78	0.79	0.94	0.91	0.86	0.77
PST (optim.)	0.86	0.79	0.89	0.89	0.95	0.93	0.91	0.85

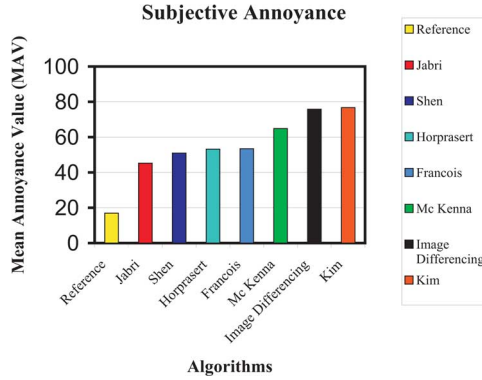


Fig. 8. $\overline{\text{MAV}}$ values obtained for each segmentation algorithm and averaged on the three tested video sequences.

narios considered. The results of the subjective experiments averaged for all the three video sequences are also reported in the last four columns. Standard methods [36] were used to analyze the judgments provided by the test subjects, to obtain the mean and to screen outliers. In order to allow a for a point-to-point comparison with the objective values obtained by the proposed metric in the result Section VI, the numerical values of the MAV are reported in Table IV. The averaged Mean Annoyance Values ($\overline{\text{MAV}}$) have been computed for each algorithm and the reference in order to provide a general overview on the segmenting performance of the described algorithms. A histogram of $\overline{\text{MAV}}$ values is also presented in Fig. 8 for the general framework discussed in the following.

In the general scenario, the subjective results show that the algorithms which on average introduce the most annoying artifacts are the Kim and **Image Differencing** algorithms. The least annoying artifacts are generated by the Jabri, Shen, and Horprasert algorithms (see Figs. 7 and 8 for visual evaluation).

The most annoying artifact is flickering usually due to noise, camera jitter, and varying illumination. It produces erroneously segmented regions (different at each frame). A high value of flickering of added regions is generated by Kim's algorithm and it is the most annoying artifact on average for the general scenario (Table IV). In fact, no matter what the size of the artifact is, if the segmentation presents temporal instabilities it will annoy the subject much more than any other spatial artifact, as suggested by both the qualitative (*interview*) and quantitative (*experimental trial*) data collected.

In the general scenario, the second most annoying artifact according to subjective experiments is that introduced by **Image Differencing** due to the large amount of holes, especially border

holes. They are perceived as the most annoying in terms of spatial errors. Holes are usually due to the algorithm's failures in differentiating the foreground regions from the background when they look very similar in color or texture or other uniformity features that the algorithm exploits to segment. Then the artifacts introduced by McKenna are rated as the third most annoying. In this case, holes are especially annoying to human observers, even if they are smaller than those introduced by the **Image Differencing** method, but still of considerable amount.

Added background is the fourth most annoying artifact and it is generated by François's algorithm. It is mostly caused by erroneously detecting moving shadows as part of the moving foreground objects. Since shadows move along with objects from which they are cast, we observed that this artifact does not annoy the human observer significantly, and is subjectively rated better than flickering or missing parts of objects in this general scenario.

The least annoying artifacts on average are introduced by the Horprasert, Jabri, and Shen algorithms. In fact, these algorithms introduce smaller amounts of artifacts compared to others. Section VI will present both the subjective and objective experimental results for specific scenarios.

V. PROPOSED EVALUATION CRITERIA

The proposed discrepancy method is defined on the basis of two types of metrics, namely the objective metric and the perceptual metric. First, the *objective metric* classifies and quantifies the deviation of the segmentation result from the reference. Second, segmentation errors are measured through the proposed objective criteria and their perception is studied and characterized by means of subjective experiments. Finally, the perception of segmentation errors is modeled and incorporated in the proposed *perceptual metric*. The novelty of our approach is to classify the different clusters of error pixels perceptual weights according to the following characteristics: whether they do or they do not modify the shape of the object and afterward their size. Border holes \mathcal{H}_b and added backgrounds \mathcal{A}_b modify the shape while inside holes \mathcal{H}_i , and added regions \mathcal{A}_r preserve the segmented object shape since they are disjoint from the correctly segmented objects (see Fig. 2).

A. Spatial Artifacts

The relative spatial error $\mathbf{S}_{A_r}(k)$, for all the j added regions at frame k , $\mathcal{A}_r^j(k)$, is obtained by simply applying

$$\mathbf{S}_{A_r}(k) = \frac{\sum_{j=1}^{N_{A_r}} |\mathcal{A}_r^j(k)|}{|n(k)|} \quad (10)$$

where $|\cdot|$ is the set cardinality operator, $n(k)$ is the sum of the reference and the result segmentation pixel sets (a normalization factor applied to all the error measures that makes the error bounded between 0 and 1), and N_{Ar} is the total number of added regions.

Similarly, for all the j holes inside the segmentation $\mathcal{H}_i^j(k)$, the relative spatial error $S_{H_i}(k)$ is given by

$$S_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |H_i^j(k)|}{|n(k)|} \quad (11)$$

where N_{Hi} is the total number of holes inside the objects. The spatial error for added background and holes on the border of the object is formulated in a different way. In fact, both kinds of errors are located around the object contours and it has to be distinguished from numerous deviations around the object boundary and a few but larger deviation [24] by adding this weighting factor D^j

$$D^j = 1 + \frac{\bar{d}^j + \sigma_d^j}{d_{\max}^j} \quad (12)$$

where d represents the distance values² of error pixels from the correct object contour. The mean \bar{d} and the standard deviation σ_d of d are calculated and are then normalized by the maximal diameter d_{\max} of the reference object to which the cluster of errors belongs to. The maximal diameter is computed by taking the maximum of all the distances of any two points on the reference object contour. By combining this last (12) and (10), we obtain, for the border artifacts, the corrected relative spatial error $S_{Ab}(k)$ for j added backgrounds

$$S_{Ab}(k) = \frac{\sum_{j=1}^{N_{Ab}} D_{Ab}^j \cdot |A_b^j(k)|}{|n(k)|} \quad (13)$$

similarly for j holes on the border $\mathcal{H}_b^j(k)$, the relative spatial error $S_{H_b}(k)$ is

$$S_{H_b}(k) = \frac{\sum_{j=1}^{N_{Hb}} D_{Hb}^j \cdot |H_b^j(k)|}{|n(k)|} \quad (14)$$

B. Temporal Artifacts

The most subjectively disturbing effect is the temporal incoherence of an estimated sequence of object masks. We studied temporal incoherence in segmented video sequences 60 frames long at 12 fps. In video segmentation, an artifact often varies its characteristics through time. A non-smooth change of any spatial error deteriorates the perceived quality. As already mentioned, the temporal artifact caused by an abrupt variation of the spatial errors between consecutive frames is called *flickering*. To take this phenomenon into account in the objective metric, a

measure of flickering is introduced, $\mathbf{F}(k)$ that can be computed for each kind of artifact $\Lambda = [A_r, A_b, \mathcal{H}_i, \mathcal{H}_b]$ as follows:

$$\mathbf{F}_\Lambda(k) = \frac{\text{abs}(|\Lambda(k)| - |\Lambda(k-1)|)}{|\Lambda(k)| + |\Lambda(k-1)|} \quad (15)$$

where $\text{abs}()$ is the absolute value operator. The difference of artifact amounts between two consecutive frames is normalized by the sum of the amount of this artifact in the current frame k and the previous frame $k-1$. In this equation, if the error disappears/appears suddenly, it is evenly penalized by the normalization since it causes an annoyance for the human observer due to the unexpected change in the segmentation quality. By doing so, the *surprise* effect [40] can also be accounted for in the metric. This effect is meant to amplify the changes in the spatial accuracy. To model this effect, (15) is combined to the relative spatial artifact measures to construct an objective spatio-temporal error measure $\mathbf{ST}(k)$ for each artifact. This takes into account not only the quality but also the stability of the results

$$\mathbf{ST}_\Lambda(k) = S_\Lambda(k) \cdot \frac{1 + \mathbf{F}_\Lambda(k)}{2}. \quad (16)$$

During informal interviewing, we discovered that flickering is the most annoying artifact. We then tested this artifact and introduced a synthetically varying amount of spatial errors at different periods and frequency. Finally, we were able to have a rough estimate of how the flicker of a spatial error influences the overall spatial temporal error. Combining the spatial artifact measure to the flickering measure as in (16) allows us to divide by 2 the perceptual annoyance of the spatial error in the case of no flickering effect present ($\mathbf{F}_\Lambda(k) = 0$). This means that the spatial error weight is half in the case that it does not flicker and the formula reflects what we observed through subjective experiments. On the other hand, if the flickering is at its maximum value ($\mathbf{F}_\Lambda(k) = 1$), the spatial error is considered in all its strength and in summary the spatial temporal error measure is larger in proportion to the flickering of the spatial error considered.

In modeling, the relation between instantaneous and overall quality [41], we can identify two other phenomena related to the temporal context, namely the *human memory* effect and the *expectation* effect. The human memory effect is related to the fact that after a while the human gets used to a certain visual quality thus judging it more acceptable if it persists long enough. In subjective experiments on coded video sequences [42], these effects have been studied. Fig. 9(a) shows the characteristics of the weighting functions for taking into account the effects of human memory. The first gradient is called the beginning effect of human memory (it lasts around 50 frames) and presents higher values at the first frames. The second gradient shows short-term human memory effect and indicates when subjects give more importance to the last frame quality, which they overall remember more.

With our subjective experiments, we aim to find the weighting function for 60 frame long video sequences. Our test videos were only 5 s long (60 frames) and thus not long enough to cause short term memory in the human observers. On the other hand, since they were short videos, we experienced a phenomenon called *expectation* effect. By expectation we mean that a good

²For distance computation, 8-connectivity has been used.

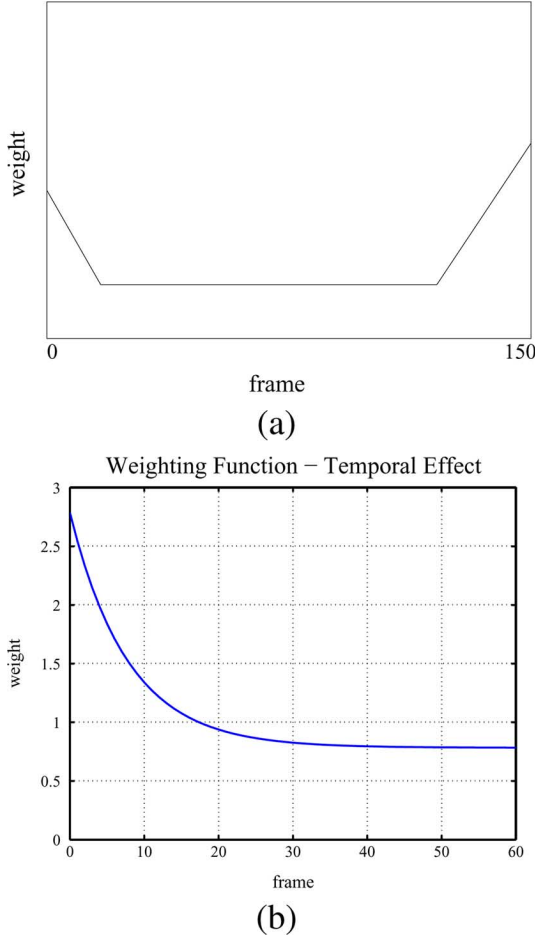


Fig. 9. Experiment on *expectation* effect. (a) Weighted function considering human memory in video quality evaluation proposed in [42]. (b) Proposed temporal weighting function taking into consideration the Expectation effect.

segmentation at the beginning could create a good overall impression on assessing the overall quality of the sequence under test, and vice-versa. To model this effect, the overall objective spatio-temporal metric \mathbf{ST} is formulated as follows:

$$\mathbf{ST}_\Lambda = \frac{1}{K} \sum_{k=1}^K w_t(k) \mathbf{ST}_\Lambda(k) \quad (17)$$

where the temporal weights $w_t(k)$ that model the *expectation* effect have been empirically defined by means of subjective tests as

$$w_t(k) = \left(\alpha \cdot e^{\frac{k-30}{\beta}} + \gamma \right) \quad (18)$$

with $\alpha = 0.02$, $\beta = 7.8$, $\gamma = 0.0078$, k is the frame. This equation represents the fitting curve we obtained from subjective experiments [14] and depicted in Fig. 9(b). As it can be noticed the judgements given during the first few frames weight much more than those given in the last frames (i.e., *expectation* effect).

C. Perceptual Objective Metric

In [16], a detailed description of the *synthetic* artifacts used to study and characterize the perception of the spatial and temporal

artifacts previously described can be found. In the following, a brief description of the parameters obtained for the perceptual metric is given and in the next section, the proposed metric is tested on *real* artifacts. The \mathbf{ST} values of each artifact metrics were plotted versus the values of MAV and the best fitting psychometric curves were found [16] to describe the human perception of errors. Four psychometric curves were derived through subjective experiments, one for each artifact, to obtain four *perceptual artifact metrics*: \mathbf{PST}_Λ . The best fitting function for each artifact was the Weibull function W . Thus, the perceptual artifact metrics are described by

$$W(x, S, k) = 1 - e^{-(Sx)^k} \text{ where } x = \mathbf{ST}_\Lambda$$

$$\mathbf{PST}_\Lambda = W(\mathbf{ST}_\Lambda, S, k) \quad (19)$$

where the parameters S and k have been obtained in [16] for the general scenario case with synthetic artifacts: $S = 0.014$, $k = 0.304$ for \mathbf{PST}_{A_s} ; $S = 0.026$, $k = 0.653$ for \mathbf{PST}_{A_b} ; $S = 0.331$, $k = 0.2339$ for \mathbf{PST}_{H_i} ; $S = 0.771$, $k = 0.641$ for \mathbf{PST}_{H_b} . In the following, the details of the subjective experiments where the parameters S and k have been obtained are summarized.

The annoyance of the **added region** artifacts was studied by varying its amount on a total of 75 sequences. Moreover, different positions and shapes of added region artifacts were tested to check if they are perceived the same way. To test this hypothesis, we used the statistical F test. In this experiment, 28 non-expert users were asked to perform the annoyance task. The subjective experiments showed that the added region annoyance perception is not influenced by the shape or position of the artifact but only by its size (see rows 1 and 2 of Table V). In the **holes** experiment, there were two goals. The first goal was to test the two objective metrics, one proposed for inside holes [see (11)] and the second for border holes [see (14)]. The second goal was to determine the psychometric annoyance functions for the two kinds of synthetic artifacts. Finally, we studied whether the annoyance caused by a boundary hole could be worse than for an inside hole (for large holes). In this experiment, 28 non-expert users were asked to perform the annoyance task on 48 test sequences. This subjective experiment indicated that both the kind and the size of the hole should be jointly taken into account and not only the distance when an objective metric is proposed. In the objective metrics proposed in the literature, holes are only considered in terms of uncorrelated sets of pixels and of their distances from the reference boundary of the object [15], [22]. With this experiment, it was proved that a cluster of error pixels should be distinguished and their characteristics should be thoroughly studied instead of considering each error pixel individually. The methods reported in [15], [22], and [27] claim that as we move away from the border, holes become more annoying, but we proved that this depends also on the type and the size of the hole [16]. Furthermore, two positions of inside holes have also been tested: one further than the other to the object borders. Hence, the statistical F-test has been used to investigate whether the perceived annoyance of these two positions could be described with two different fitting curves. As reported in row 3 of Table V, the F value shows that the same curve can be used to fit both positions for inside holes. This validates the simple characterization made about inside holes without considering the distance of the inside hole from the border of the

TABLE VII
OBJECTIVE METRIC VALUES OBTAINED FOR EACH SEGMENTATION ALGORITHM FOR ALL THE TEST VIDEO SEQUENCES IN GENERIC, COMPRESSION, SURVEILLANCE, AND MIXED REALITY FRAMEWORKS

Alg.	'Group'				'Hall monitor'				'Highway'				<i>PST</i>			
	Gen.	Compr.	Mix.	Sur.	Gen.	Compr.	Mix.	Sur.	Gen.	Compr.	Mix.	Sur.	Gen.	Compr.	Mix.	Sur.
ref.	2.96	6.84	2.95	3.45	3.67	6.42	3.56	4.28	3.67	3.78	5.78	7.28	3.43	5.68	4.09	4.09
Jab.	21.59	14.86	21.59	25.38	26.31	31.36	26.31	45.56	23.84	16.24	23.84	36.36	29.31	20.82	23.91	35.76
Hor.	31.76	43.64	31.76	36.26	31.35	40.23	31.35	35.28	20.48	20.59	20.48	23.45	27.36	34.82	27.86	31.66
Shen	28.78	40.98	35.82	24.56	35.89	63.76	35.91	35.65	24.81	37.83	24.81	33.27	29.82	47.52	32.18	31.16
Fran.	40.84	46.86	40.84	45.28	43.87	74.36	43.87	56.27	29.19	35.80	29.19	28.78	37.96	53.00	37.96	43.44
Kim	28.98	41.67	28.98	87.59	43.42	54.13	43.42	60.23	35.13	44.44	35.13	48.78	35.84	46.76	35.84	65.86
MkK.	42.73	69.18	42.73	71.23	56.86	68.26	39.41	67.34	31.12	54.66	31.12	52.71	43.57	64.03	37.75	63.76
Im.D.	46.64	92.33	46.64	43.76	60.00	62.84	28.78	53.87	36.76	50.40	36.76	43.21	47.8	68.52	37.39	46.94

ground truth [see (14) and (11)]. To further confirm the hypothesis that a distinction between inside holes and border holes has to be made, we applied the F -test on these two sets of data to see if a unique fitting curve can interpolate both kinds of artifacts (see row 4 of Table V). The F -value in this case is equal to 5.01, which is above the threshold of $F(2, 44)$ equal to 3.21. This means that an overall fitting curve is not sufficient to describe both phenomena, so two distinct metrics, one for holes on the border and one for inside holes, \mathbf{PST}_{H_b} and \mathbf{PST}_{H_i} , were proposed. These psychophysical experiments showed that inside holes for small sizes are more annoying than holes on the border, but on the other hand by increasing their size, border holes become more annoying than inside holes as the shape of the object becomes less recognizable.

The performance of the proposed objective metric for **added background** [see (13)] was tested on five dilated masks plus 16 test sequences with different amount of added background concentrated in some parts of the object boundaries. For the video *Hall monitor*, five new segmented video sequences were created by varying the number of dilations of correctly segmented video sequences from one dilation to eight dilations. There were eight subjects in this experiment. For the second test, with large amounts of added background concentrated in only some regions, 31 subjects judged the 16 test sequences. The experimental results showed that the added background measure of (13) matches the human annoyance perception both when the artifact is uniformly distributed along the object boundaries and when it is concentrated in some parts of the object boundaries. Finally, the overall perceptual metric is given by the combination of all the four types of artifacts. A simple linear combination of artifacts was found [16] to be the best estimate of the total annoyance

$$\mathbf{PST} = a \cdot \mathbf{PST}_{A_r} + b \cdot \mathbf{PST}_{A_b} + c \cdot \mathbf{PST}_{H_i} + d \cdot \mathbf{PST}_{H_b}. \quad (20)$$

The perceptual weights were found by means of subjective experiments [16] on combined synthetic artifacts: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$.

VI. EXPERIMENTAL RESULTS

In this section, two different issues are investigated. First, the performance of the proposed perceptual metric, \mathbf{PST} , *with no parameter optimization* is analyzed and compared to the state-of-the-art metrics in Section VI-A. Moreover, the results of the metric are used to discuss the performance of the selected

segmentation algorithms according to the different scenarios in order to provide general guidelines for choosing the best performing algorithm. Second, the parameters of the novel metric, \mathbf{PST} , *with parameter optimization* are obtained according to specific applications and the perceptual artifacts' weights are discussed in Section VI-B.

Before discussing the results, three issues should be clarified. First, all the subjective experiments on synthetic artifacts were carried out on the following segmented video sequences: "Group," "Hall," "Highway," and "Coastguard" (a MPEG sequence) or a subset of them. "Coastguard" has a moving background. We implemented only static-background/foreground segmentation methods. Therefore, the experiments with real artifacts always used the video sequences with static background: "Group," "Hall," and "Highway." In order to be able to generalize the results, the video content presented both "outdoor" and "indoor" scenes and we obtained the averaged results on the three video sequences in order to give some general guidelines for any video content. We also provide some *ad hoc* guidelines for specific video content suggesting which segmentation algorithm performs better for indoor scenes ("Group" and "Hall") and outdoor ones ("Highway"). This discussion is useful for helping in choosing a segmentation algorithm for the specific content/application at hand.

Second, it should be noted that the proposed perceptual metric parameters have been derived on the basis of subjective experiments on *synthetic* artifacts. By testing the metric on *real* segmentation algorithms, we want to show its excellent performance and reliability also in the case of real artifacts. These results confirm that the synthetic segmentation artifacts generated were similar to those produced by the state-of-the-art segmentation methods and the metric can be used to evaluate any real segmentation algorithm.

Third, it has to be taken into account that the proposed metric has been optimized for *synthetic* artifacts and then tested on *real* segmentation while the state-of-the-art metrics used for comparison have not been optimized.

A. Non-Optimized \mathbf{PST} Metric

The performance of the proposed \mathbf{PST} metric is analyzed in terms of correlation coefficients with the obtained subjective MAV values in Table IV. Since we did not optimize the parameters for the state-of-the-art metrics, MPEG, wqm, and mqm, in order to perform a fair comparison, in this section we kept the

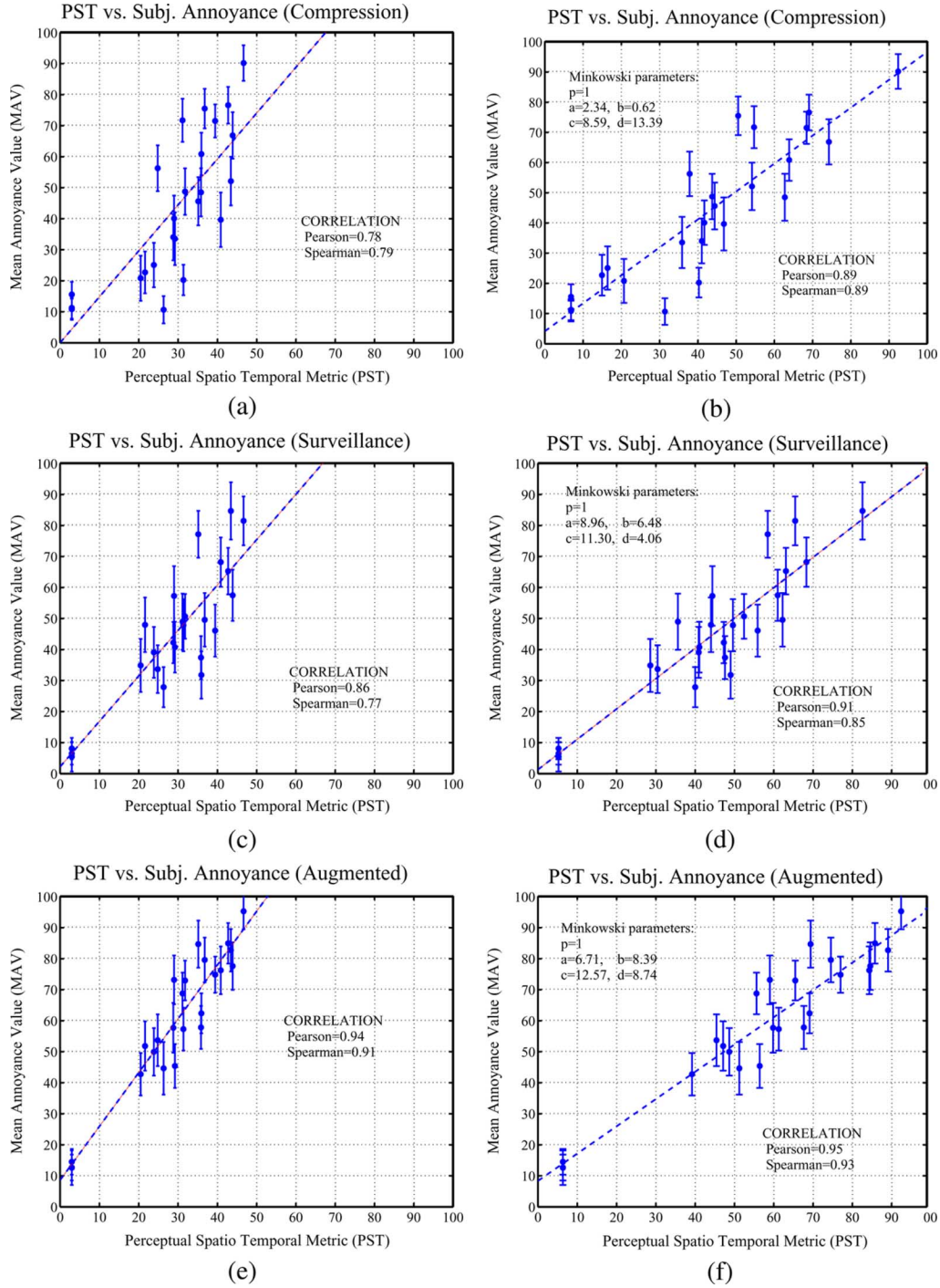


Fig. 10. Objective Metric **PST** versus Subjective Scores (MAV) and correlation coefficients for different segmentation applications: with no optimized (a), (c), (e), and optimized (b), (d), (f) Minkowski parameters.

a, b, c, d parameters of (20) equal to those obtained for the synthetic artifacts ($a = 2.86, b = 4.50, c = 4.77, d = 5.82$). These results are shown in rows 1–4 of Table VI. The linear correlation coefficient of Pearson and the nonlinear (rank) correlation coefficient of Spearman are calculated in order to correlate the subjective and the objective results in Table VI. The objective results for the considered segmentation algorithms have been plotted versus the subjective annoyance values for the four frameworks, namely general, compression, video surveillance, and mixed reality. The correlation coefficients for the perceptual

metric **PST** are larger (Pearson = 0.86, Spearman = 0.79) compared to the state-of-the-art metrics (MPEG metric, matching quality metric mqm, and weighted quality metric wqm) for all scenarios showing a good performance of the proposed metric. The perceptual metric automatically predicts the segmentation quality in a similar way to how human subjects perceive it (i.e., clusters of errors) and outperforms the state-of-the-art metrics, which do not include perceptual factors. However, MPEG metric outperforms wqm and mqm metrics in mixed reality and surveillance scenario and no

state-of-the-art metric performs well in the case of compression scenario. In the next section, we optimize the a, b, c, d parameters of (20) according to the application in order to have an experimental study to determine which are the most annoying artifacts (among added background, added region, border and inside holes) for specific applications.

Since the final goal for an objective metric is to help in choosing the best performing algorithm on a given set of data, the performance of the state-of-the-art segmentation algorithms are discussed on the basis of the MAV values reported in Table IV and **PST** metric results (without optimization) reported in Table VII. If the performance of the segmentation algorithms are considered in the general case, the best one in both subjective (Table IV) and objective (Table VII) evaluation is given by Jabri for “Hall” and “Group.” In fact, the generated confidence maps and the hysteresis thresholding method which integrates neighbor pixels is more capable than other methods to distinguish homogeneous regions. For the “Highway,” the best performance is achieved by Horprasert in which the distortions for brightness and chromaticity obtained from background modeling give a bigger range to classify only the relevant object pixels in the current frame. **Image Differencing** and Kim give the worst results due to under-segmentation and over-segmentation depending on the threshold sensitivity and the incorrect contour filling of Kim. It is important to notice that **PST** metric results in Table VII take nonzero values for the reference segmentation since the result have been normalized in order to take the same range of the subjective MAV values.

1) *Compression Scenario:* In the video compression case, overall Jabri was estimated as the best performing algorithm as for the general scenario. In fact, even if this algorithm introduces some added background and added regions, they are not so bothersome to the user in this specific application: they are not compressed as well as the rest of the object and unlike the background. **Image Differencing** and McKenna shows the worst cases since this last method is not able to deal with similar colors in the background and foreground causing inside and border holes.

2) *Mixed Reality Scenario:* In the mixed reality case, overall Jabri was still the best performing segmentation algorithm. This is due to the fact that almost no shape changes are caused by this segmentation. In fact, only few added regions are present and they do not bother the human viewers since they pay attention to the moving objects. Francois and **Image Differencing** show again the worst case since it produces a lot of inside and border holes (see Fig. 7) that allow to see the virtual background beneath the objects.

3) *Surveillance Scenario:* In video surveillance case, the biggest annoyance weights are given to added regions and inside holes. This can be explained by the fact that human viewers in the surveillance scenario pay attention to misdetected or over-detected objects that could lead to false alarms (in the case of erroneous detection of background parts as moving objects) and missed alarms (in the case of misdetection of moving objects). If the performance of the segmentation algorithms are considered in detail for the surveillance case, the best one in both subjective and objective evaluation is given by Shen. This

is due to the fact that almost no false alarms or missed alarms are caused by this segmentation. In fact, neither added regions nor missing objects are ever produced. Only few border holes and added backgrounds are present due to the integration of the motion information and a more sophisticated classification part. Kim gives the worst results due to under-segmentation and over-segmentation depending on the threshold sensitivity and the incorrect contour filling.

B. Optimized PST Metric

Our evaluation metric has been proposed for general purpose segmentation with an ideal segmentation at hand. It is important, when evaluating the performance of an algorithm, to have *a priori* knowledge on the specific application the segmentation is addressing. A novelty in the proposed metric is that the a, b, c, d parameters in (20) can be easily adjusted depending on applications by performing a Levenberg–Marquardt method for nonlinear least-squares data fitting using the subjective MAVs. Thus, on the basis of the subjective experiment, the best metric parameters have also been computed by maximizing the correlation coefficients (Pearson and Spearman) in the specific scenarios as shown in the last row of Table VI. The scatter plots in Fig. 10 show how optimized **PST** improves the correlation coefficients compared to the **PST** non-optimized and how it fits the annoyance scale [0–100] better. For all applications, the difference in perception of the four artifacts has been so quantified through the parameters a, b, c, d . The optimized parameters are reported in Fig. 10 for each scenario. The error bars in the figure represent the standard deviation of the subjective MAVs for the different subjects.

1) *Compression Scenario:* In the compression scenario, the optimized weights obtained for added regions and background ($a = 2.34, b = 0.62$) are very small compared to those for inside and border holes ($c = 8.59, d = 13.39$). In fact, in this application we have preserved the quality of the segmented objects and compressed the background. Therefore, the parts of the object that have been erroneously segmented as part of the background have been compressed and annoy the subjects more than having segmentation artifacts like added region or background that has not been compressed.

2) *Mixed Reality Scenario:* In the mixed reality scenario, the weights obtained for added background ($b = 8.31$), inside ($c = 12.57$) and border holes ($d = 8.74$) are larger than those for added background (6.71). In fact, every artifact that changes the shape or allows for seeing the virtual background beneath the real objects causes a lot of annoyance in the subjects who are focusing their attention on the virtual story or the interactive game.

3) *Surveillance Scenario:* In the surveillance application, the biggest weights are given to added regions and inside holes ($a = 8.96$ and $c = 11.30$) and the smallest to added background and border holes ($b = 6.48$ and $d = 4.06$). This can be explained by the fact that human viewers in the surveillance scenario pay attention to misdetected (inside holes) or over-detected (added region) objects that could lead to dangerous situations of false alarms (in the case of erroneous detection of

background parts as foreground objects) and missed alarms (in the case of misdetection of foreground objects).

VII. CONCLUSION

A perceptually driven objective metric for segmentation quality evaluation has been proposed on the basis of psychophysical experiments on synthetic artifacts. A study on real artifacts produced by typical video object segmentation algorithms has been carried out to test the proposed perceptual metric. To the best of our knowledge, a comparison among different state-of-the-art video object segmentation systems has received little attention by the image processing community so far, as well as the study of their performances for different applications. Seven state-of-the-art foreground/background segmentation algorithms were chosen as typical and analyzed both objectively and subjectively. First, a classification of the real artifacts introduced is provided according to subjective perception. Second, a perceptual objective metric able to predict the subjective quality as perceived by human viewers has been proposed. The results show both the better performance of such a metric compared against the usually adopted MPEG and the wqm, mqm metrics and its adaptability to take into consideration different segmentation applications. The optimal perceptual parameters have been found for specific segmentation applications: video compression, video surveillance, and mixed reality.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of this paper for their valuable and constructive review comments and to J. Klamkin for proofreading this manuscript. They would also like to thank M. Carli and G. Arrigoni for running some experiments, M. Karaman for generating some of the segmentation masks, X. Marichal and J. Nascimento for their test code.

REFERENCES

- [1] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*, 2nd ed. Tampa, FL: Thomson, 1999.
- [2] T. N. Pappas, J. Chen, and D. DePalov, "Learning perception," *OE Mag.*, vol. 5, pp. 18–20, Oct. 2005.
- [3] F. C. Pereira and T. Ebrahimi, *The MPEG-4 Book*. Upper Saddle River, NJ: Prentice-Hall, 2002.
- [4] C. Kim and J. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 122–129, Feb. 2002.
- [5] A. R. J. François and G. G. Medioni, "Adaptive color background modeling for real-time segmentation of video streams," *Int. Imag. Sci., Syst., Technol.*, pp. 227–232, 1999.
- [6] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. IEEE ICCV Frame Rate Workshop*, 1999, pp. 1–19.
- [7] J. Shen, "Motion detection in color image sequence and shadow elimination," in *Vis. Commun. Image Process.*, Jan. 2004, pp. 731–740.
- [8] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and location of people in video images using adaptive fusion of color and edge information," in *Int. Conf. Pattern Recognition (ICPR)*, Sep. 2000, pp. 627–630.
- [9] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Understanding*, vol. 80, pp. 42–56, 2000.
- [10] M. Farias, "No-reference and reduced reference video quality metrics: New contributions," Ph.D. dissertation, Univ. of Santa Barbara, CA, Aug. 2004.
- [11] S. E. Maxwell and H. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, 1999.
- [12] E. D. Gelasca, T. Ebrahimi, M. Farias, and S. Mitra, "Impact of topology changes in video segmentation evaluation," in *Proc. IEEE CVPR Workshop Image Anal. Multimedia Interactive Services*, Apr. 2004, CD-ROM.
- [13] E. D. Gelasca, T. Ebrahimi, M. Farias, M. Carli, and S. Mitra, "Towards perceptually driven segmentation evaluation metrics," in *Proc. IEEE CVPR Workshop (Percept. Org. Comput. Vis.)*, Jun. 2004, p. .
- [14] E. D. Gelasca, T. Ebrahimi, M. Farias, C. Marco, and S. Mitra, "Annoyance of spatio-temporal artifacts in segmentation quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2004, pp. 345–348.
- [15] A. Cavallaro, E. D. Gelasca, and T. Ebrahimi, "Objective evaluation of segmentation quality using spatio-temporal context," in *Proc. IEEE Int. Conf. Image Process.*, Rochester, NY, Sep. 22–25, 2002, pp. 301–304.
- [16] E. Drelie, "Full-reference objective quality metrics for video watermarking, video segmentation and 3D model watermarking," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2005.
- [17] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer, "A robust visual method for assessing the relative performance of edge detection algorithms," *Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 12, pp. 1338–1359, Dec. 1997.
- [18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. ICCV-01*, Vancouver, BC, Canada, Jul. 7–14, 2001, vol. 2, pp. 416–425.
- [19] "Compare Your Segmentation Algorithm to the Cost 211 qam," [Online]. Available: <http://www.iva.cs.tut.fi/COST211/Call/Call.htm>
- [20] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 186–200, Feb. 2003.
- [21] K. McKoen, R. Navarro-Prieto, E. Durucan, B. Duc, F. Ziliani, and T. Ebrahimi, "Evaluation of segmentation methods for surveillance applications," in *Proc. EUSIPCO*, Sep. 2000, pp. 1045–1048.
- [22] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1092–1103, Aug. 2004.
- [23] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," in *Proc. 43rd MPEG Meeting*, Tokyo, Japan, 1998, ISO/IECJTC1/SC29/WG11 M3448.
- [24] R. Mech and F. Marques, "Objective evaluation criteria for 2-D shape estimation results of moving objects," *EURASIP J. Appl. Signal Process.*, vol. 4, pp. 401–409, Apr. 2002.
- [25] C. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X Eur. Signal Process. Conf.*, Tampere, Finland, 2000, vol. 2, pp. 917–920.
- [26] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.
- [27] X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. X Eur. Signal Process. Conf.*, Tampere, Finland, 2000, pp. 2139–2196.
- [28] J. Nascimento and J. S. Marques, "New performance evaluation metrics for object detection algorithms," in *Proc. 6th Int. Workshop Perform. Eval. Tracking Surveill. (PETS, ECCV)*, Prague, Czech Republic, May 2004, pp. 7–14.
- [29] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. Joint IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2003, pp. 125–132.
- [30] R. J. Oliveira, P. C. Ribeiro, J. S. Marques, and J. M. Lemos, "A video system for urban surveillance: Function integration and evaluation," in *Proc. Int. Workshop Image Anal. Multimedia Interactive Syst.*, 2004, CD-ROM.
- [31] F. Oberti, E. Stringa, and G. Vernazza, "Performance evaluation criterion for characterizing video surveillance systems," *Real Time Imaging*, vol. 7, pp. 457–471, 2001.
- [32] A. Senior, A. Hampaput, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. 2nd IEEE Workshop Perform. Eval. Tracking Surveill.*, 2000, CD-ROM.
- [33] "IEEE and 2005 and Winter Vision, Performance Evaluation of Tracking and Surveillance (PETS)." [Online]. Available: <http://pets2005.visualsurveillance.org/>
- [34] "EU CAVIAR Project Caviar Test Case Scenarios." [Online]. Available: <http://homepages.inf.ed.ac.uk/rbf/CAVIAR> ,
- [35] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

- [36] "Subjective video quality assessment methods for multimedia applications recommendation," Int. Telecomm. Union, Geneva, Switzerland, 1996.
- [37] "Methodology for subjective assessment of the quality of television pictures recommendation," Int. Telecomm. Union, Geneva, Switzerland, BT.500-1, 2002, .
- [38] P. Milgram and H. Colquhoun, "A taxonomy of real and virtual world display integration," in *Mixed Reality, Merging Real and Virtual Worlds*, Y. Otha and H. Tamura, Eds. New York: Ohmsha-Springer, 1999.
- [39] X. Marichal, B. Macq, D. Douxchamps, and T. Umeda, "and art.live consortium the ART.LIVE architecture for mixed reality," in *Proc. VRIC*, Laval, France, Jun. 2002, pp. 19–21.
- [40] J. W. Senders, "Distribution of visual attention in static and dynamic displays," *SPIE, Human Vis. Electron. Imaging II*, vol. 3016, pp. 186–194, Feb. 1997.
- [41] R. Hamberg and H. de Ridder, "Time-varying image quality: Modeling the relation between instantaneous and overall quality," *SMPTE J.*, vol. 108, pp. 802–811, 1999.
- [42] Y. Inazumi, Y. Horita, K. Kotani, and T. Murai, "Quality evaluation method considering time transition of coded quality," in *Proc. ICIP*, Oct. 24–28, 1999, vol. 4, pp. 338–342.



Elisa Drelie Gelasca (M'08) received the M.S. degree from the University of Trieste, Trieste, Italy, in 2001, and the Ph.D. degree from the Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2005.

She served as a Research Consultant at Genista SA, a high-tech start-up company in the field of multimedia quality metrics. In June 2006, she joined the Vision Research Laboratory, University of California at Santa Barbara (UCSB), where she is currently a member of the Center of BioImage and Informatics.

Her research interests are in the area of image processing, with special emphasis on bioimage analysis, benchmarking, duplicate detection, perceptual quality assessment of video segmentation, watermarked three-dimensional objects, and video sequences.



Touradj Ebrahimi (M'08) received the M.Sc. and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1989 and 1992, respectively.

In 1993, he was a Research Engineer at the Corporate Research Laboratories, Sony Corporation, Tokyo, Japan, where he conducted research on advanced video compression techniques for storage applications. In 1994, he served as a Research Consultant at AT&T Bell Laboratories working on very low bitrate video coding. He is currently a

Professor at EPFL heading its Multimedia Signal Processing Group. He is also an Adjunct Professor with the Center of Quantifiable Quality of Service at the Norwegian University of Science and Technology (NTNU). He has initiated more than two dozen national, European, and international cooperation projects with leading companies and research institutes around the world. He is also the head of the Swiss delegation to MPEG, JPEG, and SC29, and acts as the Chairman of Advisory Group on Management in SC29. He is a cofounder of Genista SA, a high-tech start-up company in the field of multimedia quality metrics. In 2002, he founded Emitall SA, start-up active in the area of media security and surveillance. In 2005, he founded EMITALL Surveillance SA, a start-up active in the field of privacy and protection. He is a member of the Scientific Advisory Board of various start-up and established companies in the general field of information technology. He has served as Scientific Expert and Evaluator for Research Funding Agencies such as those of European Commission, the Greek Ministry of Development, the Austrian National Foundation for Scientific Research, the Portuguese Science Foundation, as well as a number of venture capital companies active in the field of information technologies and communication systems. His research interests include still, moving, and 3-D image processing and coding, visual information security (rights protection, watermarking, authentication, data integrity, steganography), new media, and human–computer interfaces (smart vision, brain computer interface). He is the author or the coauthor of more than 200 research publications and holds 14 patents.

Prof. Ebrahimi has been the recipient of various distinctions and awards, such as the IEEE and Swiss National ASE Award, the SNF-PROFILE Grant for Advanced Researchers, Four ISO-Certificates for key contributions to MPEG-4 and JPEG 2000, and the Best Paper Award of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS. He became a Fellow of the International Society for Optical Engineering (SPIE) in 2003. He is or has been Associate Editor with various IEEE, SPIE, and EURASIP journals, such as the *IEEE Signal Processing Magazine*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, *EURASIP Image Communication Journal*, *EURASIP Journal of Applied Signal Processing*, and *SPIE Optical Engineering Magazine*. He is a member of SPIE, ACM, and IS&T.